Westmere-EX: A 20 thread server CPU

Dheemanth Nagaraj, Sailesh Kottapalli
Westmere-EX Architecture

Acknowledgements:
Westmere-EX Team

Legal Disclaimer

- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Nehalem, Merom, Boxboro, Millbrook, Penryn, Westmere, Sandy Bridge and other code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user
- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.
- Intel, Intel Inside, Intel Core, Intel Xeon, Intel Core2 and the Intel logo are trademarks of Intel Corporation in the United States and other countries.
- *Other names and brands may be claimed as the property of others.
-

Risk Factors

This presentation contains forward-looking statements that involve a number of risks and uncertainties. These statements do not reflect the potential impact of any mergers, acquisitions, divestitures, investments or other similar transactions that may be completed in the future. The information presented is accurate only as of today's date and will not be updated. In addition to any factors discussed in the presentation, the important factors that could cause actual results to differ materially include the following: Demand could be different from Intel's expectations due to factors including changes in business and economic conditions, including conditions in the credit market that could affect consumer confidence; customer acceptance of Intel's and competitors' products; changes in customer order patterns, including order cancellations; and changes in the level of inventory at customers. Intel's results could be affected by the timing of closing of acquisitions and divestitures. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of new Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel's response to such actions; Intel's ability to respond quickly to technological developments and to incorporate new features into its products; and the availability of sufficient supply of components from suppliers to meet demand. The gross margin percentage could vary significantly from expectations based on changes in revenue levels; product mix and pricing; capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; excess or obsolete inventory; manufacturing yields; changes in unit costs; impairments of long-lived assets, including manufacturing, assembly/test and intangible assets; and the timing and execution of the manufacturing ramp and associated costs, including start-up costs. Expenses, particularly certain marketing and compensation expenses, vary depending on the level of demand for Intel's products, the level of revenue and profits, and impairments of long-lived assets. Intel is in the midst of a structure and efficiency program that is resulting in several actions that could have an impact on expected expense levels and gross margin. Intel's results could be impacted by adverse economic, social, political and physical/infrastructure conditions in the countries in which Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the report on Form 10-Q for the quarter ended June 27, 2009.

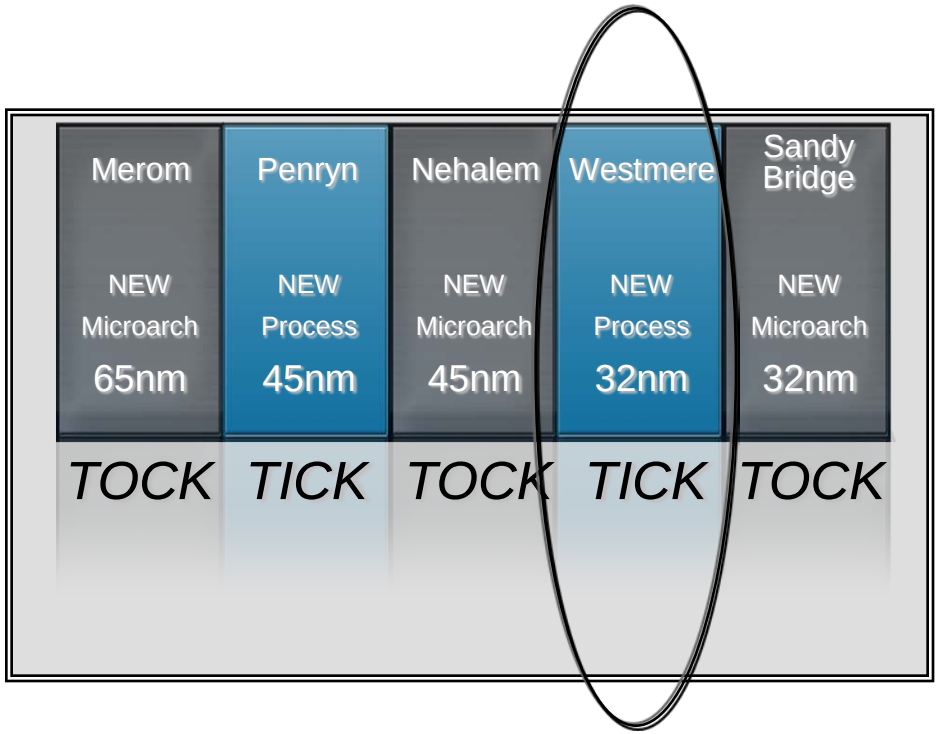
Today's Talk

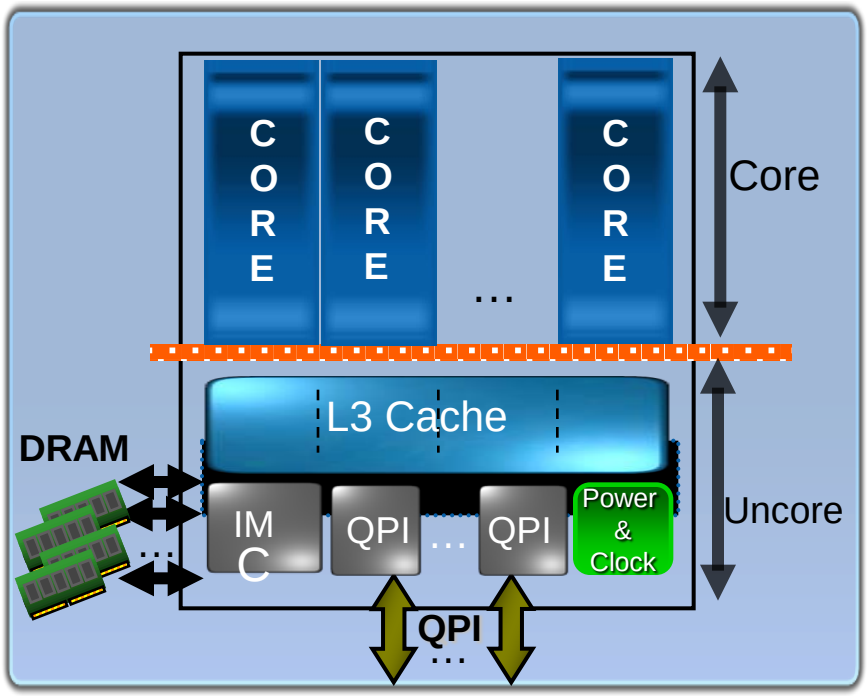
- Westmere-EX (WSM-EX) Processor discussion
  - Baseline architecture
  - Focus on enhancements along key vectors: Performance, Power, RAS, Security and Virtualization
- Things we are not disclosing today
  - Product clock Speeds
  - Product performance/power
  - Overall Intel Server Roadmap

Agenda

- Westmere-EX Architecture Overview
- Balanced Performance Scaling
  - Core/Cache scaling
  - Bandwidth, Protocol enhancements
- Power Management
  - Core/Package idle state support
  - Package idle sub-states
    - Low Power memory links, memory self-refresh
    - Macro level clock gating
- Memory RAS
  - Double Device Data Correct (DDDC)
- Security and Virtualization
- Conclusions

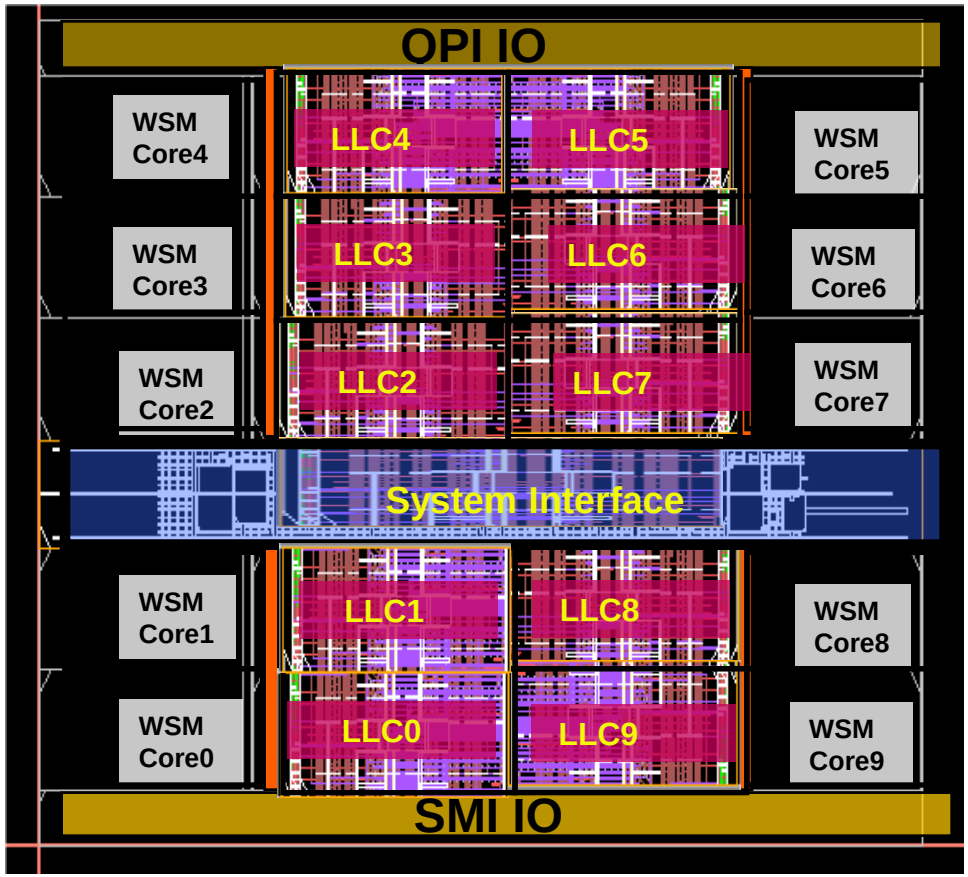
Tick-Tock Converged Core Development Model





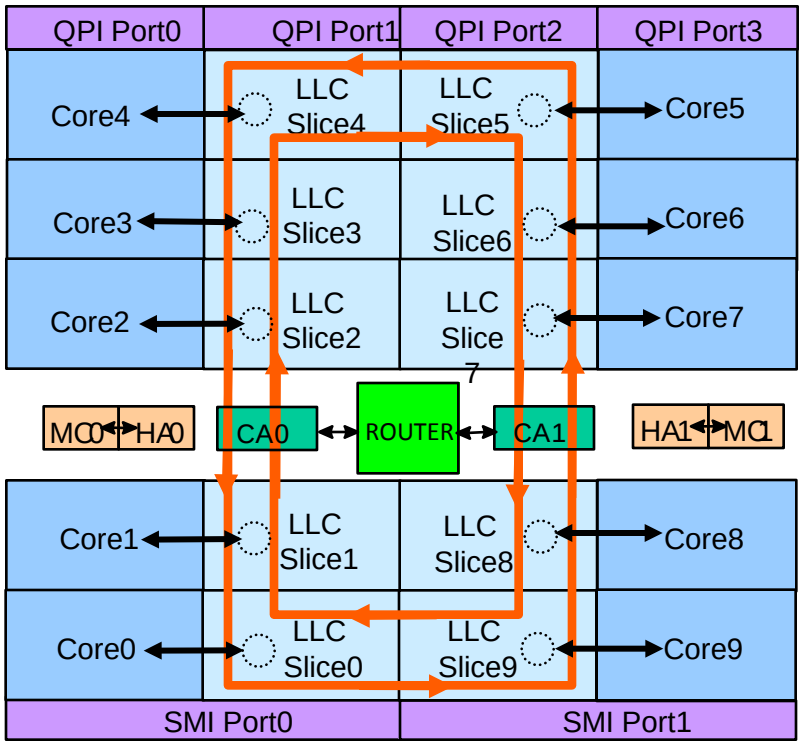
- Converged WSM core for first generation of 32nm client and server CPUs
  - Server specific feature support and reliability requirements incorporated in the core
- Uncore, core count differentiates product segment specific CPUs

WSM-EX CPU



- Refresh to Boxboro-EX platform; Socket compatible with Xeon® 7500
- 10 WSM cores, 20 threads; monolithic die
- 10 slice shared Last Level cache (L3)
- 2 integrated memory controllers
- 4 Quick Path Interconnect (QPI) system interconnect links
- Scalable Memory Interconnect (SMI) with support for up to 8 DDR channels
- Supports 2, 4 and 8 socket in glueless configs and larger systems using Node Controller (NC)
- Intel 32nm process technology

Micro-Architecture Overview - 1



HA - Home Agent (Coherence Agent)

CA – Caching Agent Hub

MC – Memory Controller

Refer Xeon® 7500 HotChips21 presentation

www.hotchips.org/archives/hc21/2_mon/HC21.24.100.ServerSystemsI-Epub/HC21.24.122-Kottapalli-Intel-NHM-EX.pdf

- Distributed 10 slice, shared LLC (L3 cache)
  - 10 way Physical Address hashing to avoid hotspots
  - 5 parallel LLC access requests per clock
  - 32B (half cache-line) wide data-path
- Bi-directional scalable ring interconnect
  - Ring stops hook up a core/LLC slice, CA to the ring
  - LLC miss traffic funneled through CA0/CA1
  - CA0 proxies slice0-4 and CA1 proxies slice5-9
  - BW scales with added core/LLC ring stops
- Structural imbalances addressed through slot reservation
  - Ring protocol provides priority to a message on the ring over a new message
    - Simplifies protocol but exacerbates imbalances among ring stops in funneling traffic through CA.
    - Outer ring stops can continuously pump CA bound messages and starve inner stops
  - Starvation could lead to excessive ring bounce and BW tailbacks (lower BW at higher traffic injection)
  - Starvation resolved by reserving slots that can only be used by inner ring stops

Micro-architecture Overview - 2

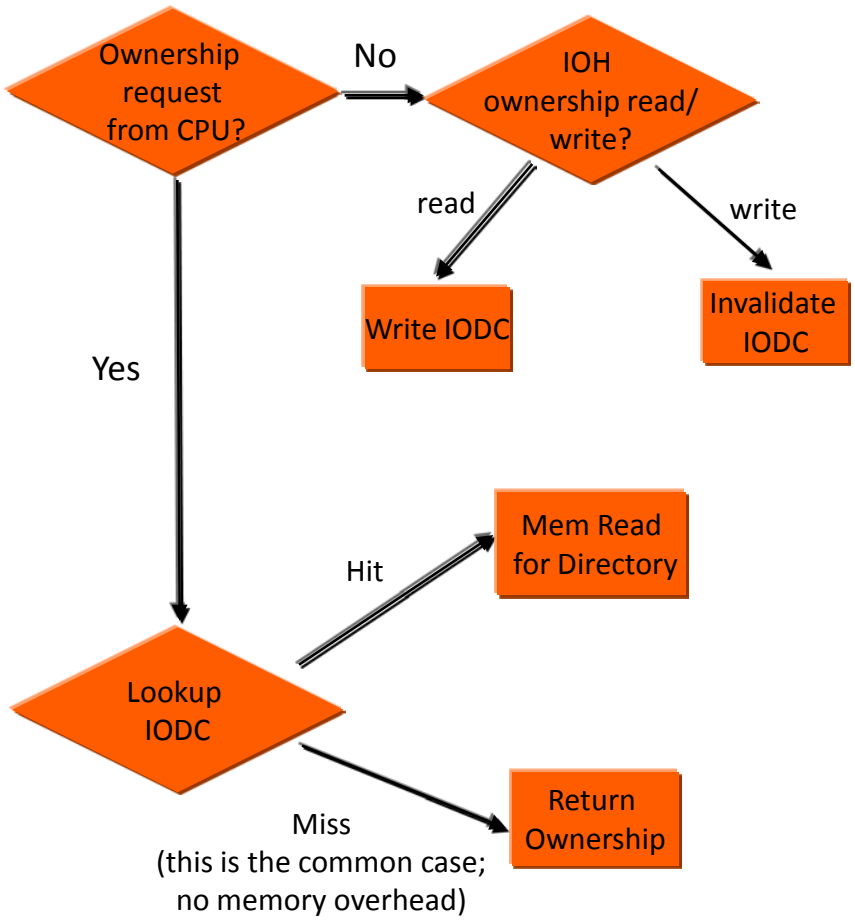
- Up to 120 outstanding requests supported
  - Supports mapping all requests to local socket memory for NUMA optimized workloads
- QPI Source snoopy protocol with in-memory IO-Hub (IOH) ownership tracking
  - CA snoops peer sockets on every read request
  - HA snoops IOH based on directory
  - 256 pre-allocated requests per HA (512 per socket)
- Up to 96 outstanding requests at memory across 2 Memory Controllers (MCs)
  - MC supports Out-of-Order scheduling across non-conflicting requests
  - Scheduling done at rank granularity; Per rank blackout counters enforce DDR spec timing adherence
  - Pair of lock-stepped Scalable Memory Interconnect Links per MC
    - Lock stepped channels enables advanced memory RAS

Balanced Performance Scaling

- Core/Cache scaling through a modular architecture, scalable interconnect
  - cache sized to mitigate memory BW demand increase from added cores
- Raw, application BW improved through micro-arch and scheduler changes
  - Number of outstanding requests increased
    - CA (48 -> 60), MC (32 -> 48)
  - Scheduling optimized for Scalable Memory Buffer expansion topology
    - Per rank blackout timer counters to track DRAM timing
    - Flexibility to differentiate Same Rank, Same DIMM different rank and Different DIMM turn-around timings
  - CA outstanding memory request capping policy in the QPI pre-allocated request scheme augmented to improve NUMA performance
  - IO Directory cache to aid applications with non-temporal stores (covered later)
- Protocol enhanced for Directory Assisted Snoopy flow
- Micro-architecture driven Operating voltage optimization
  - IO digital logic moved from high speed link clock domain to the lower frequency uncore domain while maintaining delivered BW
  - Lowers Operating voltage; Not constrained by high speed physical layer logic

IOH Directory Cache (IODC)

- IOH ownership is tracked through in-memory directory
  - Reduces snoop BW requirement on the IOH
- IODC Improves delivered BW for applications with Non-Temporal stores (HPC apps)
  - Non-Temporal hints allow a store directly to memory without a fetch
  - Stores need to be write combinable to a full cacheline write.
  - Caching Agent (CA) spawns an ownership request + memory writeback
- Without IODC, Home Agent (HA) will issue a memory read to ascertain IOH ownership information from in-memory directory
  - The directory lookup read is wasted BW from the application perspective
- With IODC, memory read is avoided by caching the IOH ownership information
  - The memory BW matches the BW delivered to the application

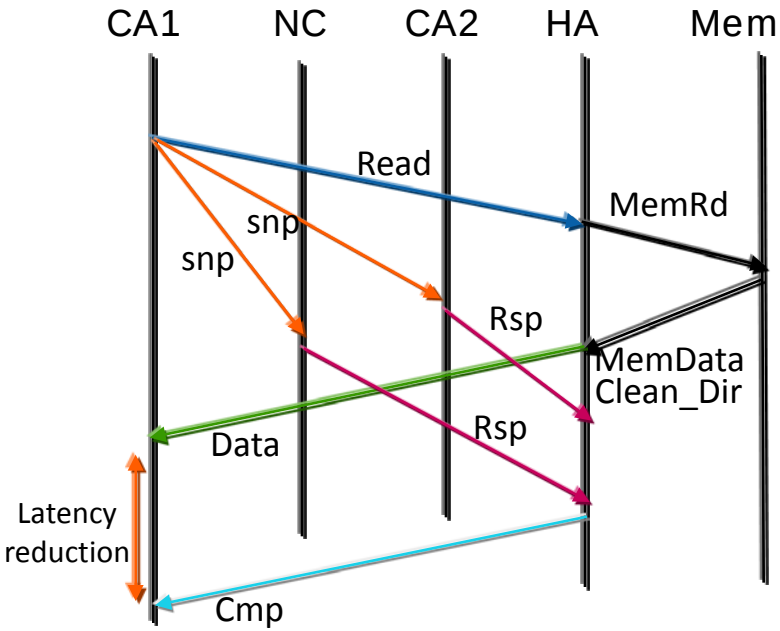


IODC Flow Diagram

Directory Assisted Snoopy (DAS)

- Targets local socket memory latency reduction for snoop bound topologies
  - 8-socket glueless and some Node-Controller based platforms
- In-memory directory augmented to track remote socket cacheline ownership (R-state indicates remote ownership)
  - Data returned to local requestor without waiting on snoop responses on a clean directory
  - R state Directory tracking adds some bandwidth overhead; mitigated for NUMA optimized workloads

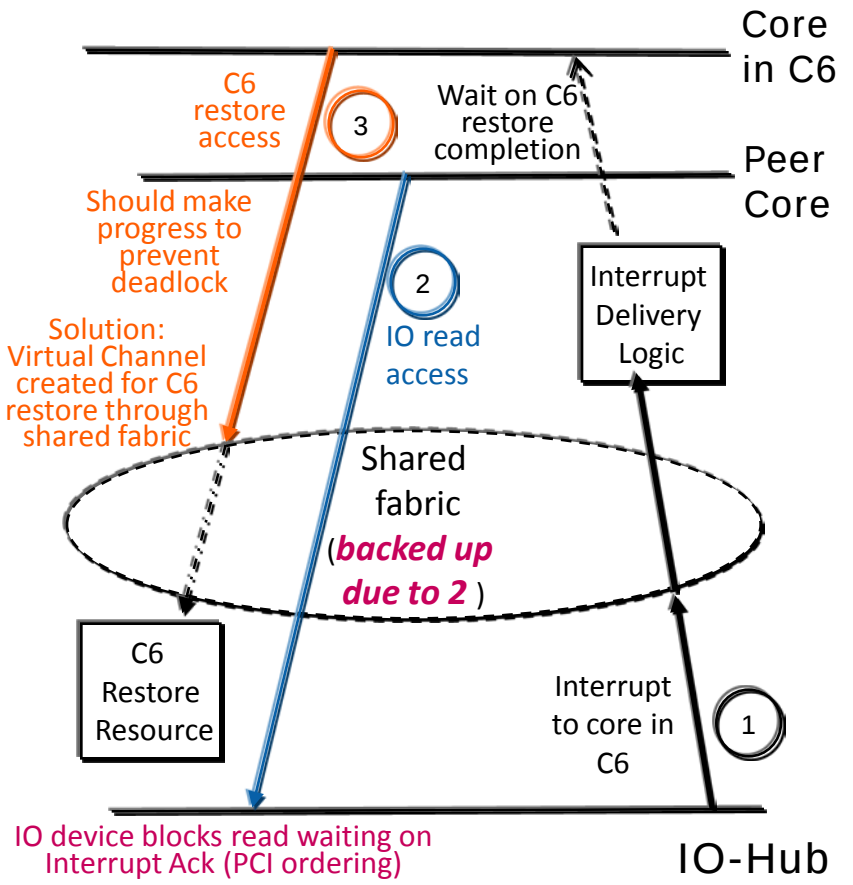
DAS Flow for Local Read with Clean Directory



Power Management – Core power down (C6 state)

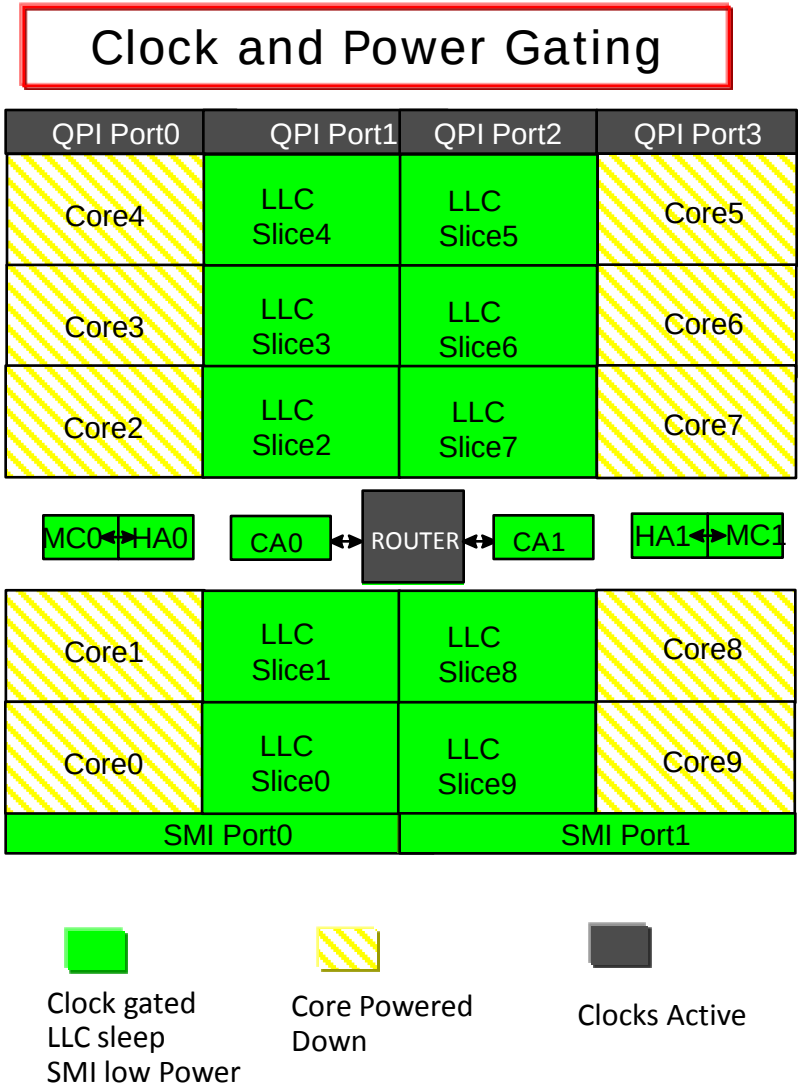
- Core C6 supported to eliminate core leakage power when OS deems threads to be idle
  - Per core power gates controlled through power control unit in the uncore
  - Additionally, improves dynamic frequency boost characteristics of active core
- Intricate deadlocks resolved with C6 entry/exit flows
  - Interplay with QPI protocol message dependencies, PCI ordering rules, lock and other global flows
- Package idle state (PC6) entered through negotiation with other agents in the platform
  - All cores in C6 triggers negotiation
  - Favorable response indicates platform idle state => memory access latency resulting from further package low power actions tolerable
  - Negotiation allows early wake-up indication to peer agents well before traffic is generated from woken-up cores

C6 exit deadlock due to PCI ordering interaction
IORead (2) waits on **Interrupt (1)** Ack – PCI ordering
Interrupt (1) waits on **C6 restore(3)** - core wake-up
C6 restore (3) is backed-up behind **IoRead(2)** => deadlock



Power Management – PC6 Sub-states

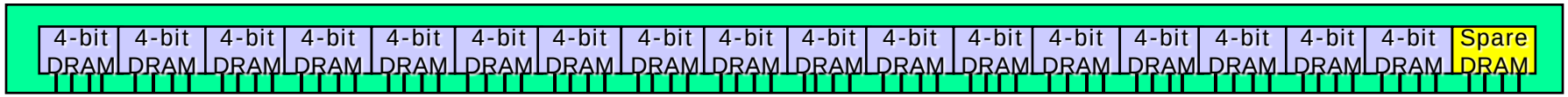
- Entry into Package C6 state allows additional power saving actions
- Low Power SMI link state with memory put in self refresh
  - CPU firmware initiated entry; autonomous exit on memory request
  - Exit latency optimized to ensure DMA request latency does not back pressure network packets
- Macro level Clock gating on bulk of uncore logic
  - Gating done at Regional clock buffer
  - In-band traffic from QPI links blocked at router input port until clocks are un-gated
  - Out of band events (pin based interrupts) routed to power management firmware to generate a wake-up for the clocks



Memory Availability – Double Device Data Correct (DDDC)

- Enables recovery from up to 2 DRAM chip failures per X4 memory rank
  - Recovery from additional bit flip (DDDC + 1) supported
  - Separate trials for X8 and for X4 correction supported
    - X4 device contributes 16 bits per read; Requires 32 bits of redundancy for detect and correct
    - X8 device contributes 32 bits per read; Requires 64 bits of redundancy for detect and correct
    - Correction at X4 granularity enables optimal use of available redundancy
- Implemented through a combination of micro-code, hardware, BIOS

Rank with no device fails



Rank with 2 device fails (spare + in-line correction to recover)

Security and Virtualization

- Advanced Encryption Standard-New Instruction (AES-NI) ISA extension for cryptographic acceleration
  - 6 new instructions (encryption, decryption, key generation and carry less add)
  - All 3 NIST specified Key sizes can be supported
  - Refer AES-NI public documentation: http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/
- VT-x3 real mode addressing and 1GB page supported
  - Guest operation in real mode removes the performance overhead and complexity of an emulator
- Private per thread memory provided in the CPU to cache the entire Virtual Memory Control Structure (VMCS)
  - Improves VM switch latency

Westmere-EX Summary

- Compelling refresh to Boxboro-EX Platform in the Intel 32nm process generation
- Focus on balanced feature set across Scalable Performance, Power management, RAS, Security and Virtualization
- Directory look-up overheads and source snoop latency performance bottlenecks addressed through micro-architectural innovations
- Brings state of the art idle power management features to the EX space
- Builds on the WSM core security and virtualization hooks

Glossary

AES-NI: Advanced Encryption Standard – New Instructions
Boxboro: IO-Hub chipset associated with Xeon® 7500 and Westmere-EX CPU
C-States: Refers to processor Idle states ranging from halt to power-down
C6 : Core Power Down Idle state
CA: Caching Agent
EX: Expandable server segment. Nomenclature replaces MP
HA: Home Agent (coherency controller)
iMC: Integrated Memory Controller
IODC: IO Directory Cache
IOH: IO Hub chipset
MC: Memory Controller
PA: Physical Address
PC6: Package C6
QPI: Quick Path Interconnect System interconnect link.
SMI: Scalable Memory Interconnect
Socket: CPU die
Uncore: Logic on the CPU die excluding the code. Includes LLC, System Interface logic
VMCS: Virtual Memory Control Structure